Heavy-tailed statistics of cortical representational drift are advantageous for stabilised downstream readouts

Charles Micou[1] and Timothy O'Leary[1]

[1]Department of Engineering, University of Cambridge, United Kingdom

25th September 2024

Abstract

Neural representations of familiar environments and mastered tasks continue to change despite no further refinements to task performance or encoding efficiency. Downstream brain regions that depend on a steady supply of information from a neural population subject to this representational drift face a challenge: they must stabilise their readout using only statistical regularities in neural activity. Recent studies have described how representational drift induces deterioration in the accuracy of fixed decoders. Here, we highlight that while a variety of underlying statistics of drift at the level of individual neurons can produce comparable deterioration of a fixed decoder, these different statistics result in dramatically different deterioration rates in adaptive decoders. We describe an adaptive decoding strategy that can read out from a model drifting population long after fixed decoders degrade completely, and demonstrate that in simulated models this strategy is better-suited to heavy-tailed drift statistics, in which individual neurons make sudden and large changes in tuning. We investigate the extent to which these advantageous statistics manifest in experimental in-vivo measurements of drift by considering two existing and well-studied datasets that observe drift in the posterior parietal cortex and the visual cortex. We find preliminary support for sudden jumps in neural tuning that would allow a downstream observer to more easily distinguish changes in representation from noise. These observations are a step towards refining the larger picture of mechanisms underpinning the robustness of information transfer between brain regions that can function in spite of changes in representation driven both by drift and by the learning of new information.

Introduction

Individual neurons within populations have no direct access to a ground truth. Instead, cortical populations must build up a representation of the external world through statistical regularities in upstream activity. However, these statistical regularities in upstream activity are non-stationary: they change over time in a way that is not mean-reverting. While these non-stationary changes can be the result of an upstream population itself forming new representations, a wealth of observations find neural representations of familiar environments and mastered tasks changing without any obvious external evidence of learning or changes in behaviour [1]. The nature of representational drift induces a gradual degradation in the accuracy of a fixed readout, eventually rendering that readout no better than chance.

In order to extract a steady stream of information from a drifting population, downstream consumers of that population, whether they be artificial neural decoders or neural populations, require a mechanism

to adapt their readouts. How the brain implements this mechanism remains an intriguing puzzle, with existing work describing both feedback mechanisms for externally driven correction to readouts [2] and biologically plausible self-correcting neural codes [3]. A core challenge for these mechanisms is the difficulty of distinguishing nonstationary changes in tuning from statistical fluctuations: how can they determine that a new activity pattern corresponds to true changes in tuning, rather than merely a noisy observation?

Emerging opinions present drift, or at least some subcomponent of drift, as a direct consequence of the repeated application of the learning rules implemented by neural populations [4, 5]. We often describe drift as a learning-driven walk exploring a degenerate parameter-space [6, 7]. Indeed, stochastic gradient descent-based algorithms for training artificial neural networks produce not only drift-like explorations of the parameter-space, but also phenomena with physiological analogues: networks trained at high learning rates, with smaller batch sizes, or using drop-out drift faster [8, 9], much like how in-vivo drift rates appear modulated by frequency of stimulus exposure or brain region-specific plasticity [10, 5]. Unmodified stochastic gradient-descent is a technique that, by construction, produces incremental changes in the parameter-space of a model. This conveniently produces the Brownian motion-like drift we see at the level of entire populations, with each representation being somewhat similar to the last [11].

However, our central argument in this work is the observation that there are many statistical models of drift at the level of parameters–the tuning of individual neurons–that can generate gradual drift in representation-space, that not all of these are functionally equivalent, and that gradual drift in the tuning parameters of individual neurons may in fact be disadvantageous. Crucially, in the context of downstream populations producing stable readouts on the basis of statistical regularities in upstream activity, there is an advantage in distributions of tuning parameter changes that are 'heavy-tailed'. When drift is mediated by few neurons making large changes in their tuning, rather than all neurons making many small ones, it becomes easier to distinguish nonstationary changes in representation from mere statistical fluctuations, which facilitates the adaptation of readouts.

To support this argument, we first present a first-principles computational model of a method to stably extract an accurate signal from a drifting population long after a fixed readout would have degraded beyond the point of usefulness. We describe this method in terms of an 'adaptive decoder': an algorithm that can infer changes in how stimuli are encoded in a population. We emphasise that our formulation is not intended to be a faithful description of an algorithm implemented by the physiology of the brain, but instead as a best-case normative example of such a mechanism. We use this adaptive decoder, which is constrained to rely only on statistical regularities in activity to update its parameters, to illustrate fundamental limitations in the extraction of underlying representational changes from noisy observations, and how these limitations are mitigated by heavy-tailed statistics. We then demonstrate that the advantages of these heavy-tailed statistics present a design trade-off: compensating for drift without these heavy tails requires either larger, more redundant neural populations, or the rates of drift must be slower. Finally, we explore whether biology makes use of these advantageous statistics by considering two existing and previously studied in-vivo datasets that exhibit representational drift. We characterise the distributions of tuning changes within these populations and find that they also demonstrate a preference for heavy-tailed statistics, especially under high rates of drift. However, before presenting these results, we briefly outline the background for experimental observations of drift and some of their more promising explanations.

Background

Drift is distinguished from simple noisy observation by its non-stationary statistics and resulting gradual degradation of fixed readouts. Drift is distinguished from the other form of non-stationary changes in neural populations, learning, by a constant representation fidelity: drifting representations do not represent the same variables more accurately (e.g. by increasing potential decoder accuracy), more efficiently (e.g. by using fewer neurons), or by using different overall encoding strategies (e.g. moving from a sparse to a dense encoding).

Drift has been observed in a variety of brain regions: in the hippocampus [12, 13, 14], the visual cortex [15, 16], the posterior parietal cortex [17], and the olfactory cortex [10]. Identification of drift requires robust tracking of individual neurons over the course of days to weeks, and so these studies predominantly use Calcium imaging techniques in mouse or rat models. While fixed decoders that use electrophysiology to record neural activity also degrade in performance over time, including human datasets and with state-of-the-art probes for closed-loop Brain-Machine Interface applications [18], it can be challenging to confidently distinguish effects of slight electrode movement or changing in recording conditions from measurements of drift.

Even with ideal imaging registration, difficulties remain in establishing which components of observed drift can be attributed to true changes in representation when subtle changes in behaviour between experimental sessions can manifest in the activity of populations [19, 20]. This difficulty is exacerbated by the possibility that internal states such as engagement or hunger can alter representations [21, 22]. Despite these challenges and potential confounds, diligent experimental work has repeatedly uncovered evidence of drift.

Measured in-vivo drift rates vary considerably. Drift is observable in mere hours to days in highly plastic regions such as the hippocampus [13], but only on much longer timescales (weeks to months) in less plastic regions such as the visual cortex [16]. This relationship suggests a strong link between learning rates and drift rates, and additional factors such as the degree of novelty [10], the amount of experience [13], or potentially even the complexity of the task or richness of environment can modulate drift rate [5].

We supplement the predictions in this work based on computational models with observations from two existing and well-studied datasets that differ from each other in task design and imaged brain region. The first dataset, originally published in Driscoll *et al.* (2017) [17], contains recordings from the posterior parietal cortex (PPC) in mice during expert performance of a delayed choice task in a virtual T-maze [23]. This work originally measured drift by considering a sparse encoding of position along the main arm of the maze, with mean neuron firing forming peaks at specific locations, and measuring the changes in peak location. Recording sessions are spaced one day apart, and drift is visible between individual sessions. The second dataset, originally published in Marks & Goard (2021) [16], contains recordings from the V1 region of the visual cortex in mice during exposure to both orientated grating stimuli and movies of naturalistic scenes. Instead of measuring drift through explicitly parameterised tuning curves, this work analysed clusters of neurons with comparable activity and used changed in membership of those clusters to estimate drift rates. This drift manifests more slowly than in the PPC dataset, and is measured in sessions spaced a week apart.

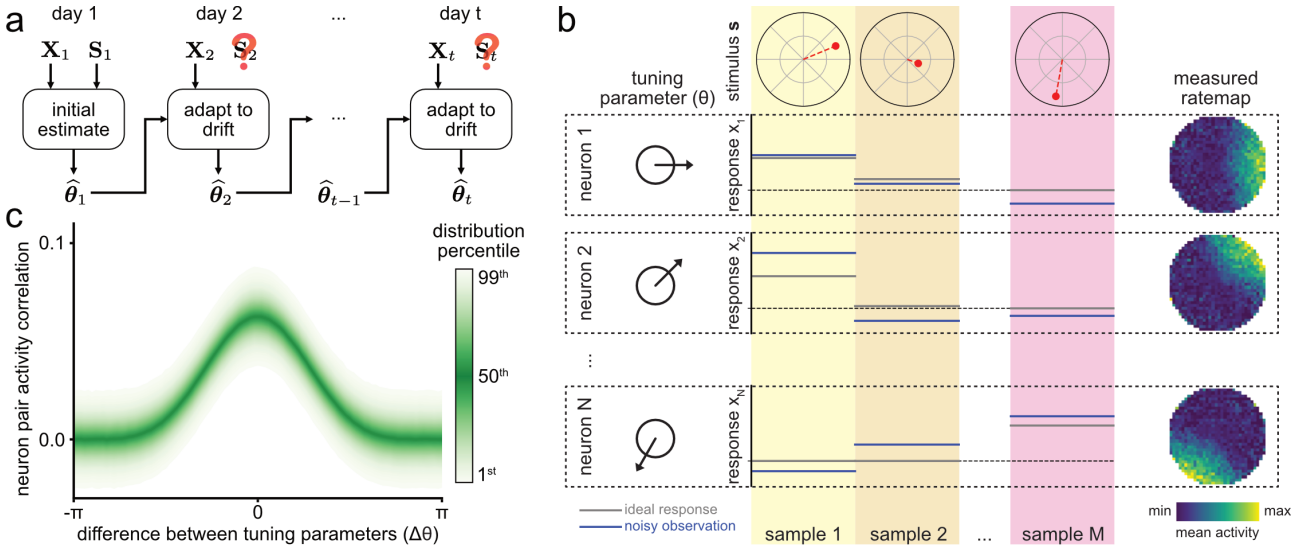

Figure 1: **(a)** Knowledge of the stimulus ground-truth from an initial experimental session is used to initialise an estimate of the tuning parameters that describe the relationship between stimulus and population activity. Subsequently, and with no further reference to the stimulus ground-truth, an adaptive decoding scheme attempts to keep these estimates accurate while the underlying tuning parameters drift. **(b)** A model of a neural population that encodes the orientation and magnitude of a stimulus. Each neuron in the population has a tuning parameter that describes its preferred orientation. The population is sampled multiple times during a session, with the activity of each neuron independently affected by noise at each sample. **(c)** Visualisation of the probability distribution for the value of the correlation between a pair of neurons as a function of the difference between their tuning parameters (simulation parameters $M = 10^4$ and $\sigma = 1.0$).

Results

Adaptive decoding via continual tuning parameter inference

The objective of an adaptive decoder is to produce an accurate readout from the population in spite of drift. Each day, the adaptive decoder updates its estimates of how individual neurons are tuned to stimuli. Outside of initialisation, the adaptive decoder cannot 'cheat' with knowledge of stimulus values: it must instead use changes in the statistical regularities of the population to infer changes in the tuning of individual neurons.

Concretely, we model a neural population of N neurons by parameterising their responses $\mathbf{x} \in \mathbb{R}^N$ to a given stimulus $\mathbf{s} \in \mathbb{R}^K$ by defining P tuning curve parameters per neuron. The complete tuning of the population is therefore described by $\boldsymbol{\theta} \in \mathbb{R}^{N \times P}$. Individual neurons, subject to independent random noise ξ, have their responses defined by an activation function a common to all neurons in the population that relates tuning parameters to activity:

$$x_i = a(\mathbf{s}, \theta_i) + \xi \tag{1}$$

On the first day of an experiment, $t = 1$, we estimate of the tuning parameters $\widehat{\boldsymbol{\theta}}_{t=1}$ of the population by taking M samples of neural activity and the stimulus ($\mathbf{X}_{t=1} \in \mathbb{R}^{N \times M}$ and $\mathbf{S}_{t=1} \in \mathbb{R}^{K \times M}$ respectively). On subsequent days $t = 2, 3, ..., T$, the tuning parameters $\boldsymbol{\theta}$ change due to drift and the ground-truth values of the stimulus are withheld from the adaptive decoder. The objective of the adaptive decoder is to maintain an accurate estimate of $\widehat{\boldsymbol{\theta}}_t$ in spite of these restrictions (Fig. 1.a). For a set of tuning parameters $\boldsymbol{\theta}$, we can use the activation function a, a prior distribution for stimulus exposure, and a model of the noise process

for ξ to construct the function g, an approximation of the likelihood of M observations of neural activity $\mathbf{X}$.

$$p(\mathbf{X}|\boldsymbol{\theta}) \approx g(\mathbf{X}, \boldsymbol{\theta}) \tag{2}$$

Naïvely, we might attempt to recover $\widehat{\boldsymbol{\theta}}_t$ from $\mathbf{X}_t$ directly by finding the parameters that maximise $g(\mathbf{X}, \boldsymbol{\theta})$. However, because the activation function a is common to all neurons, the solution-space for maximising $g(\mathbf{X}, \boldsymbol{\theta})$ has symmetries that render the solution-space degenerate. Because a describes all neurons as interchangeable, a set of parameters maximising $g(\mathbf{X}, \boldsymbol{\theta})$ could be shuffled and assigned to entirely different neurons without affecting the solution: without their history, neurons lack identity.

By considering the tuning of neurons on the previous day, $\boldsymbol{\theta}_{t-1}$, we are able to break these symmetries. We use a stochastic model of how drift perturbs the population from one day to the next, $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$. This produces a maximum a posteriori update rule for tuning in the population.

$$\widehat{\boldsymbol{\theta}}_t = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}_t) \tag{3}$$

$$\text{where} \quad \mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}_t) = \log\left(p(\mathbf{X}_t|\boldsymbol{\theta})p(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{t-1})\right) \tag{4}$$

Treating drift as a stochastic process that is independent of neuron identity and in which each neuron drifts independently from other neurons, we can express $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ in terms of a per-neuron drift prior f.

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \prod_i^N f(\theta_{t,i}, \theta_{t-1,i}) \tag{5}$$

We then combine our model for observed activity g with our description for drift f and express the daily update rule for $\widehat{\boldsymbol{\theta}}$ as:

$$\widehat{\boldsymbol{\theta}}_t = \arg\max_{\boldsymbol{\theta}} \left(\log g(\mathbf{X}_t, \boldsymbol{\theta}) + \sum_i^N \log f(\theta_i, \widehat{\theta}_{t-1,i})\right) \tag{6}$$

Statistical regularities in the activity of a simulated population

As an illustrative example of this adaptive decoding technique, we consider a simple model of a neural population that encodes the direction and the magnitude of a stimulus: a system that is reminiscent of in-vivo populations both of head-direction coding in the hippocampal formation and the response of populations in the visual cortex to grating orientations (Fig. 1.b). The population responds to a stimulus that corresponds to a point within the unit circle, defined by a magnitude and direction (i.e. $K = 2$), both of which are drawn from uniform distributions:

$$\mathbf{s} = \begin{pmatrix} r \\ \phi \end{pmatrix} \sim \begin{pmatrix} U(0,1) \\ U(0,2\pi) \end{pmatrix} \tag{7}$$

Neurons each have a single tuning parameter ($P = 1$), their preferred stimulus angle, and are subject to additive Gaussian noise.

$$x_i = a(r, \phi, \theta_i) + \xi \tag{8}$$

$$\text{where} \quad \xi \sim N(0, \sigma^2) \tag{9}$$

$$\text{and} \quad a(r, \phi, \theta_i) = \begin{cases} r\cos^2(\theta_i - \phi) & \text{if } 0 \leq \angle(\theta_i, \phi) \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Where for the convenience of dealing with the circular geometry of this example, we have defined $\angle(\psi_1, \psi_2)$ as the magnitude of the smallest angle between the angles ψ_1 and ψ_2. In this simple example, the mean decoding error in heading direction is directly equivalent to the mean error in the estimate of the tuning parameters.

To construct an expression for $g(\mathbf{X}, \boldsymbol{\theta})$, we consider pairs of neurons and the correlation of their activities within a session. Neurons with similar tuning values are more likely to be highly correlated with each other (Fig. 1.c). The correlation between neurons of a given angular separation is a probability distribution governed by two underlying random processes: the M random samples observed on any given day, and the observation noise σ perturbing those observations (Fig. S1.a).

Taking M to be sufficiently large (Fig. S1.b) that the expected value of the correlation between two neurons of an angular separation $\Delta\theta$ is well-approximated by:

$$E(\Delta\theta) = \frac{1}{6\pi}\frac{1}{16}\left(3\sin 2\Delta\theta + 2(\pi - \Delta\theta)(\cos 2\Delta\theta + 2)\right) \tag{11}$$

The distribution of the residual between the expected value and the measured value is normally distributed (Fig. S1.c). This allows a simple expression for the likelihood of the correlation a pair of neurons $\mathbf{x}_i$ and $\mathbf{x}_j$ with tuning parameters θ_i and θ_j observed M times within a session:

$$p(\mathbf{x}_i, \mathbf{x}_j | \theta_i, \theta_j) \approx h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) \tag{12}$$

$$= \frac{\sqrt{M}}{\sigma\sqrt{2\pi}} e^{-\frac{M(\mathbf{x}_i \cdot \mathbf{x}_j / M - E(\angle(\theta_i, \theta_j)))^2}{2\sigma^2}} \tag{13}$$

Treating the probabilities of pairwise correlations as independent, which is an approximation of $p(\mathbf{X}|\boldsymbol{\theta})$, we can then construct an expression for $g(\mathbf{X}, \boldsymbol{\theta})$:

$$g(\mathbf{X}, \boldsymbol{\theta}) = \prod_i^N \prod_j^N h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) \tag{14}$$

When solving the optimisation for $\widehat{\boldsymbol{\theta}}_t$, this amounts to minimising the sum of squared differences between observations of the correlations between neuron pairs and their expected values $\boldsymbol{E}$ as parameterised by $\boldsymbol{\theta}$.

Contrasting statistics of drift at the level of individual neurons

In our illustrative model the need to break symmetries using a drift prior is obvious: while $g(\mathbf{X}, \boldsymbol{\theta})$ helps determine the angular differences between neuron tuning parameters, it provides no information about the

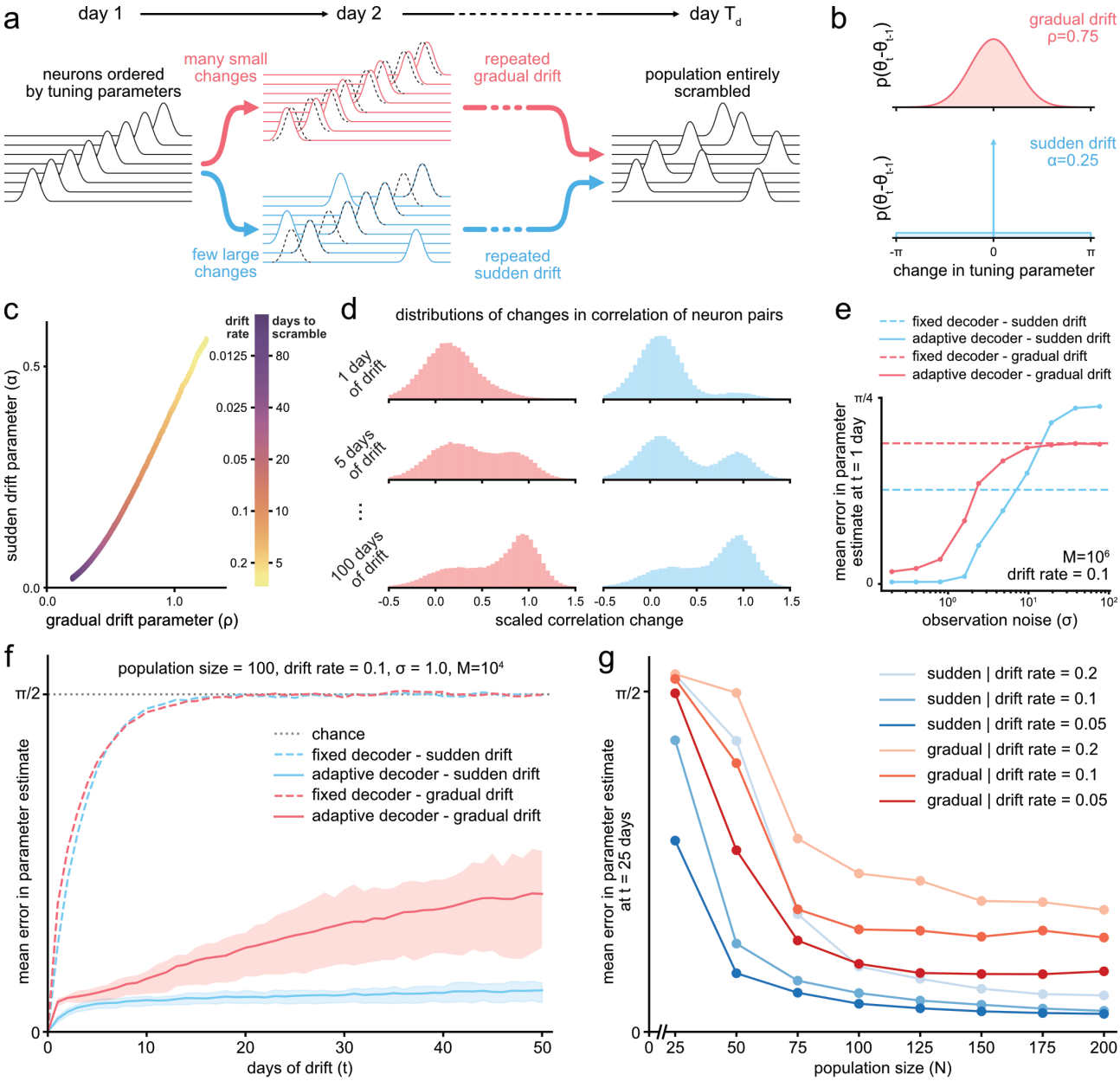

Figure 2: **(a)** Different underlying stochastic processes at the level of individual neurons, such as either gradually altering all neurons slightly or altering only a few neurons dramatically, can both produce drift. **(b)** These two example models of drift can be expressed as probability distributions describing the change in tuning parameter from one day to the next. **(c)** Equivalence between the two drift models determined by the expected number of days required to drive the error of a fixed decoder to within 5% of chance. **(d)** Histograms for the distribution of scaled changes in correlations between neuron pairs for both the sudden and gradual models of drift, shown for a relatively slow drift rate of 0.02. Only neuron pairs with a correlation of at least 0.05 on the latter day of comparison are displayed. **(e)** The resulting error in tuning parameter estimates after the first day of drift, shown as a function of the observation noise σ for both gradual and sudden drift models. **(f)** Comparison of the accuracy of parameter estimates between fixed and adaptive decoders for both sudden and gradual drift models, both using drift rates of 0.1. Shaded regions show the interquartile range over 100 simulations. **(g)** The estimation error after 25 days shown for a variety of drift rates (all sufficiently high to scramble the population before 25 days) and for a variety of population sizes. Each datapoint represents the mean of 100 simulations.

absolute orientation of any tuning parameter. Knowledge of how the population was tuned on an earlier day, as well as a drift prior describing how that tuning is likely to change, is required to anchor the system to an absolute frame of reference. Different models of drift can produce very similar degradation in the accuracy of a fixed decoder while having very different statistics at the level of individual neurons. We consider two simple but quite opposite models for drift in individual neuron tuning: a 'gradual' model, in which all neurons undergo relatively small tuning changes from one day to the next, and a 'sudden' (or 'heavy-tailed') model, in which most neurons retain their tuning while a small subset retune by making relatively large changes to their tuning (Fig. 2.a, Fig. 2.b).

In the gradual drift model, each parameter θ_i moves around the unit circle as if buffeted by Brownian motion: it is described by an Ornstein-Uhlenbeck process. The new value of a tuning parameter measured on a subsequent day is parameterised by ρ, the expected absolute distance drifted.

$$\theta_{i,t}|\theta_{i,t-1} \sim N(\theta_{t-1}, \rho^2) \tag{15}$$

$$f_{\text{gradual}}(\theta_{t,i}, \theta_{t-1,i}) = \frac{1}{\rho\sqrt{2\pi}} e^{-\frac{(\angle(\theta_{i,t},\theta_{i,t-1}))^2}{2\rho^2}} \tag{16}$$

In the sudden drift model, each neuron undergoes a spontaneous retuning instead of gradually adjusting its tuning parameter. On any given day, it either retains its previous tuning parameter or selects an entirely new value. This distribution is parameterised by the spontaneous retuning probability α.

$$f_{\text{sudden}}(\theta_{t,i}, \theta_{i,t-1}) = \begin{cases} \frac{\alpha}{2\pi} + (1-\alpha)\delta(\theta_{t,i} - \theta_{i,t-1}) \text{ for } 0 \leq \theta_{t,i} < 2\pi \\ 0 \text{ otherwise} \end{cases} \tag{17}$$

Where δ denotes the Dirac delta function.

To allow a fair comparison between sudden and gradual drift models, we define the drift rate of two models to be equivalent when the repeated application of drift scrambles the output of a fixed decoder in the same expected number of days T_d (Fig. 2.a).

$$\text{drift rate} = \frac{1}{T_d} \tag{18}$$

In the case of our illustrative example, we quantify the change in an individual neuron's tuning between day 1 and day t as the angular distance between tuning parameters $\angle(\theta_1, \theta_t)$. As drift progresses, the expected value of the change in tuning approaches the expected value between the initial tuning parameter θ_1 and one chosen entirely at random θ_r.

$$\lim_{t\to\infty} E[\angle(\theta_1, \theta_t)] = E[\angle(\theta_1, \theta_r)] = \frac{\pi}{2} \tag{19}$$

We describe a population as 'scrambled on day T_d' if it satisfies:

$$\frac{E[\angle(\theta_1, \theta_r)] - E[\angle(\theta_1, \theta_{T_d})]}{E[\angle(\theta_1, \theta_r)]} < \epsilon \tag{20}$$

Which allows us to generate an equivalence between the parameters ρ and α of the two drift models (Fig. 2.c). We choose $\epsilon = 0.05$, noting that the equivalence is not sensitive to this choice of parameter for small ϵ

(Fig. S2.a, Fig. S2.b).

While the two models of drift provide identical long-term behaviour, the differences in short-term, transient behaviour have the potential for material impact on our adaptive decoding strategy. A small change in tuning might easily be confused for a noise-driven fluctuation (due either to the observation noise σ or a small number of samples M), while a large change in tuning is more distinctive. To visualise this effect, we consider the distribution of changes in the correlation between pairs of neurons. We define the scaled change in correlation across two days D_1 and D_2 as:

$$\Delta C = \frac{C_{D_2} - C_{D_1}}{C_{D_2}} \quad \text{for} \quad C_{D_2} \geq C_{\min} \tag{21}$$

To aid the visualisation of the distribution we include only neurons with some small minimum correlation $C_{\min}$ on D_2, as most neurons are not correlated with each other, which introduces a positive bias to the distribution. When we compare these distributions for the sudden and gradual models of drift (Fig. 2.d), we notice that over short time separations the sudden drift model is more distinctively bimodal. A single day apart, most measured changes in pairwise correlation are driven by statistical fluctuations: the component of the distribution near $\Delta C = 0.0$. However, some neurons that were previously uncorrelated with each other have become correlated, accounting for the component of the distribution away from the origin. In contrast, under gradual drift it is more difficult to distinguish whether an individual change is due to a statistical fluctuation or a nonstationary change.

Sudden drift is advantageous for adaptive decoding

In simulations of using the adaptive decoding scheme in our illustrative model, there is an obvious discrepancy between gradual and sudden drift in the difficulty of distinguishing statistical fluctuations from nonstationary changes. On the very first day of drift, the adaptive decoder can accurately infer changes in the tuning parameters of the population for both types of drift under low observation noise σ. However, as σ increases, adaptive decoders make larger errors when inferring tuning parameters subject to gradual drift (Fig. 2.e).

Adaptive decoders remain accurate long after fixed decoders perform no better than chance. However, because the strategy for estimating parameters relies on using estimates from earlier days, errors can compound and eventually degrade adaptive decoders (Fig. 2.f).

As the number of neurons in the population increase, adaptive decoders maintain their accuracy for longer for two reasons (Fig. 2.g). First, while additional neurons requires estimating an increasing number of parameters, the introduction of those neurons provide more neuron pairs with which to narrow the distribution of $p(\mathbf{X}|\boldsymbol{\theta})$. The number of additional parameters to estimate grows with N, but the number of pairs available grows with N^2, which reduces the likelihood of poorly estimating the tuning of neurons. Second, additional neurons provide redundancy that is relevant over the course of many days of drift: it is statistically inevitable that, eventually, the tuning parameter of an individual neuron will be poorly estimated. With greater redundancy in the population, it is easier to notice an individual poor tuning parameter estimate on some later day, allowing subsequent correction of the estimate and forestalling compounding errors. Sudden drift has an advantage not only because it introduces smaller errors under noisy observations from one day to the next (Fig. 2.e), but also because it can more effectively use of redundancy to fix error compounding

(note the flattening gradient in Fig. 2.f and the better scaling with population size in Fig. 2.g), as from the perspective of the adaptive decoder a neuron with a low likelihood of its estimated tuning parameter is no different from a neuron that has spontaneously retuned.

Unsurprisingly, both gradual and sudden drift become easier to correct for under lower rates of drift, as the distribution of changes in tuning is narrower. Nevertheless, for the same population size and the same rate of drift, sudden drift is always easier to adapt to (Fig. 2.g). The learning that takes place within real neural populations may necessitate gradual changes, but supporting those gradual changes would require either larger populations or slower changes to achieve error rates equivalent to what is achievable with sudden tuning changes.

As an aside, we note that the definition for equivalence between sudden and gradual drift rates, which considers a long timescale over which the representation becomes completely scrambled, may be present an unfair disadvantage to gradual drift, whose Brownian motion-like exploration of parameter space can return to previous values. Using gradual drift to scramble the population beyond recognition in the same number of days as the sudden drift induces more error in a fixed decoder after the first day of drift (as visible in Fig. 2.e). However, even if we redefine the equivalence between sudden and gradual drift to cause identical degradation in a fixed decoder after the first day of drift (resulting in sudden drift models completely scrambling populations faster than gradual ones), we still find the same advantages of sudden of gradual drift (Fig. S2.c, S2.d).

Is in-vivo drift sudden or gradual?

Our computational results imply that neural populations with downstream readouts can facilitate drift compensation by making representational changes stark, especially in the case of fast drift rates. This would allow for populations to adjust faster or employ fewer neurons. Given the potential advantages, we expect that the statistics of in-vivo drift reflect a preference for heavy-tailed drift.

To explore whether in-drift manifests drift suddenly or gradually, we return to our previous comparison of the distribution of pairwise correlations between neurons (as in Fig. 2.d). This has the advantage of being agnostic towards which features of environments, stimuli, or tasks a given population encodes, and therefore allows us to draw crude comparisons both between different in-vivo datasets and between in-vivo datasets and in-silico models.

We reiterate that because both gradual and sudden drift will eventually scramble a population entirely, it becomes impossible to distinguish gradual from sudden drift over timescales that are long relative to the drift rate. In gradual models of drift, the correlation between a pair of neurons gradually increases or decreases as the neurons drift together or apart, while in sudden drift their correlation appears or disappears spontaneously. In the transient period in which it is possible to distinguish gradual from sudden drift, we model the distribution of these pairwise tuning changes as a mixture of two component Gaussian distributions, with one component describing noise-driven fluctuations and the other describing drift-driven changes in tuning (Fig. 3.a, Fig. 3.b). While simplistic, this mixture model allows us to qualitatively distinguish gradual from sudden drift. Under gradual drift, the component of this distribution ascribed to changes in tuning gradually separates from the component ascribed to noise. We can therefore identify gradual drift by inspecting the increase over time in the mean of component of the mixture model that corresponds to

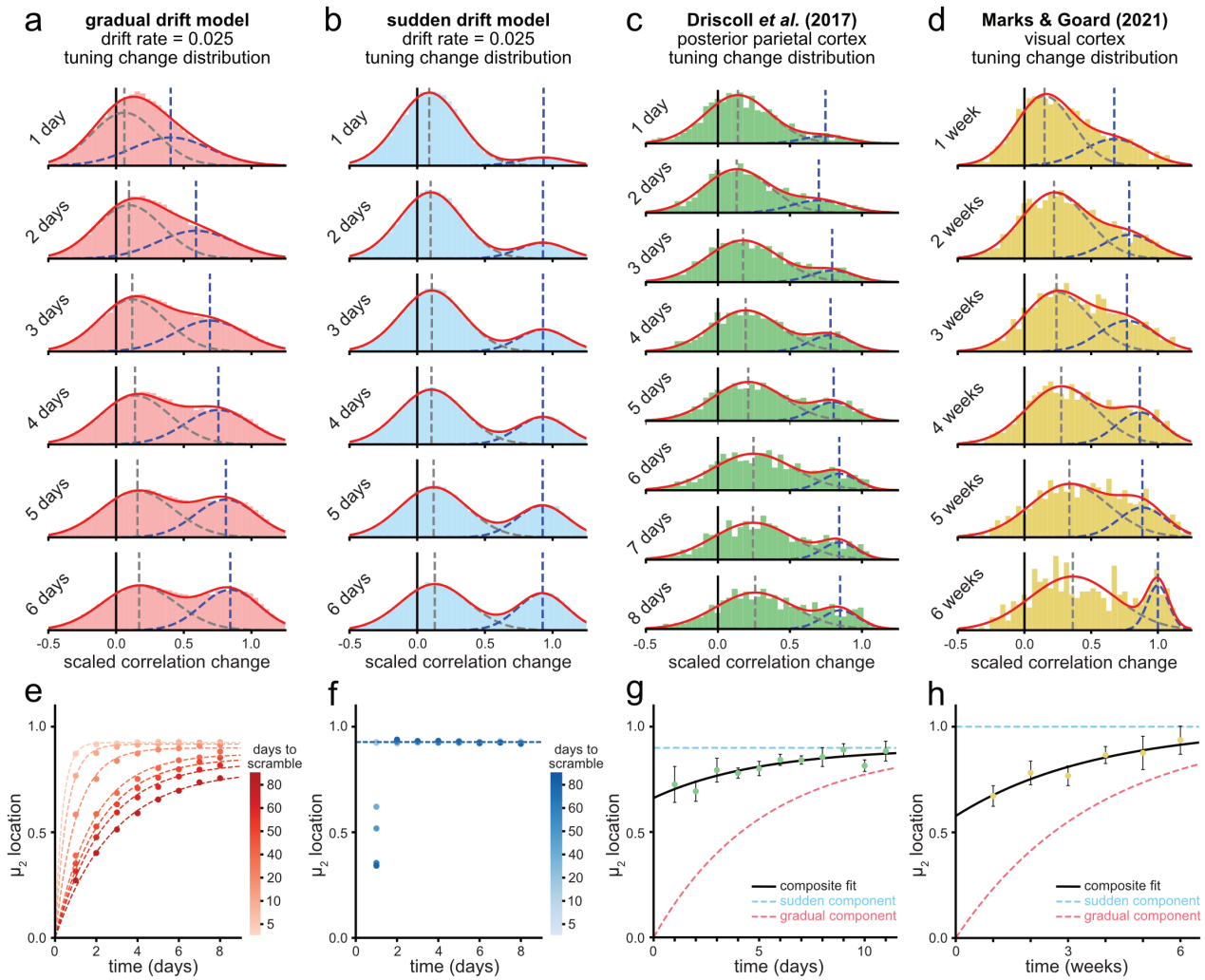

Figure 3: **(a)** Histograms showing the evolution of the distribution of changes in scaled correlations between neuron pairs in the gradual model of drift over the course of several days. The fit is a 2-component mixture of Gaussians, with each underlying component and its mean shown as dashed lines. **(b)** As in (a), but shown for the sudden model of drift at an equivalent drift rate and simple final-value fit. **(c)** As in (a), but shown for in-vivo data from the Driscoll *et al.* (2017) dataset [17]. This comprises recordings from the posterior parietal cortex during performance of a T-Maze task. **(d)** As in (a), but shown for in-vivo data from the Marks & Goard (2021) dataset [16]. This comprises recordings from the visual cortex during exposure to movies of naturalistic scenes. **(e)** Location of the upper mean (μ_2) in the mixture of Gaussians fit for the gradual drift model shown against time for a range of drift rates. The dashed line indicates a fit to the exponential approach approximation. **(f)** As in (e), but shown for the sudden model of drift and using a constant value approximation. **(g)** Location of the upper mean (μ_2) in the mixture of Gaussians fit for the Driscoll *et al.* dataset. Error bars indicate +/- SEM as estimated via bootstrap. The fit is a weighted combination of gradual and sudden drift models, with each component rendered as a dashed line. **(h)** As in (g), but shown for the Marks & Goard (2021) dataset.

nonstationary changes (μ_2), the behaviour of which can be approximated by an exponential decay towards a final value (Fig. 3.e). By contrast, under sudden drift, the two components are distinct from the outset, and there is therefore minimal change in the value of μ_2 (Fig. 3.f).

We first consider the dataset of recordings from the PPC in Driscoll *et al.* (2017) [17], where drift is observable on the timescale of days. On visual inspection, changes in the distribution of tuning do not fit neatly into the mould of either sudden or gradual drift (Fig. 3.f). While many changes appear to be mediated by large jumps in tuning, a gentle trend in the growth of μ_2 implies that some component of the drift is also mediated by more gradual changes. We fit a simple weighted linear combination of our approximations for the sudden and gradual drift models for the evolution of μ_2, and attribute relatively more of the changes in the distribution to sudden changes (Fig. 3.g).

Producing the distribution of correlation changes requires neuron pairs to remain intact between imaging sessions and to be have well-correlated activity on at least one of the sessions. The distribution in Fig. 3.f is an aggregate of all animals in the dataset. While some animals have more imaged neurons and more recording sessions (Fig. S3.a) and drift rates are not necessarily homogenous across animals, the proportion of representation of each animal within the data remains largely unchanged as a function of separation between sessions (Fig. S3.b), and therefore a change the ratios of which animals are represented is not responsible for changes in the shape of the distribution.

Next, we consider the dataset of recordings from V1 in Marks & Goard (2021) [16], where drift is much slower and observable on the timescale of weeks (Fig. 3.d). This too contains a mixture of sudden and gradual drift, but by comparison to the PPC data appears to contain a greater proportion of gradual drift (Fig. 3.h). As with the PPC data, this distribution is an aggregation over multiple animals with varying amounts of neuron pairs (Fig. S3.d and Fig. S3.d). The two datasets, which differ in rates of drift, differ also in the amount of drift which is mediated by gradual changes: the population with the slower drift rate is able to make a greater proportion of changes gradually.

Discussion

In this work, we presented an argument and computational simulations in support of the advantages of populations mediating changes in representation through stark and sudden changes in the tuning of individual neurons. We are quick to highlight that the advantages of sudden changes in neuronal tuning do not preclude the existence of gradual changes: on the contrary, we demonstrate that it is possible to adapt to such gradual changes, and that at least some component of experimentally observed drift can be attributed to these gradual processes. However, we note that to support such gradual changes, the population undergoing such changes must make concessions: it must either use a larger number of neurons and thus a higher degree of redundancy for its representations, or it must drift (or learn) more slowly, or it must rely on some external feedback system to compensate for drift.

We described a simple model process for heavy-tailed drift, a sudden retuning to a completely new parameter, to illustrate its advantages for distinguishing statistical fluctuation from change in tuning. In practice, other processes can perform this role better through distributions with even heavier tails. For example, upon retuning a neuron might be forced to adopt a new tuning parameter that is sufficiently different from its previous value. This 'active avoidance' of previous tuning values makes it even easier to

distinguish statistical fluctuation from nonstationary change.

We acknowledge that there are myriad explanations for a discrepancy in sudden versus gradual tuning adjustments between the PPC and V1. In addition to the potential for physiological differences, these are different experiments recorded under different conditions during the performance of different tasks. Nevertheless, this preliminary consideration of in-vivo data presents rudimentary agreement with the premise that a neural population can facilitate adaptation within downstream readouts by mediating representational changes through sudden changes in tuning, and that this advantage is more pronounced in situations involving higher drift rates.

The relevance of the advantages of heavy-tailed changes in tuning extends beyond representational drift. Other non-stationary changes in the encoding of information of a population, such as the incorporation of new information during learning, face the same challenge of making themselves visible through the noise of day-to-day fluctuations. If we entertain the view that some component of drift is in fact caused by the repeated application to the same data of the same continual learning rules that incorporate new information, we open up the possibility of deriving constraints on the learning rules implemented by the brain through measurements of drift. While it is experimentally challenging to attribute new behaviours or newly learned information to changes in in-vivo physiology, we speculate that–if measured–such changes would also demonstrate stark, heavy-tailed statistics.

Equally, we see future relevance of this work to in-silico learning algorithms. Artificial neural networks are prone to drift when trained using gradient descent on a feed of continuously supplied data. This drift can, in turn, cause catastrophic forgetting for samples that have remained unseen for long periods of time. While techniques exist either to restrain this drift by explicitly incorporating penalty terms in loss functions [24] or by mitigating its consequences by actively identifying and utilising subspaces of parameter-space orthogonal to previously learned items [25], we are intrigued by the possibility that a combination of network redundancy and regularisation that discourages relatively small parameter changes might provide an implicit solution. By altering the dynamics of gradient descent to explore parameter-space by focusing, rather than distributing, updates to weights in a network, exposure to new samples disturbs previously learned representations in a way that discourages a tug-of-war between samples and instead encourages implicit identification of orthogonal subspaces of parameter-space.

The two key components of the implementation of our adaptive decoder are an accurate drift prior and a description of how neurons tune to a particular kind of stimulus in a population. While we do not claim that downstream populations explicitly perform Bayesian optimisation to account for changes in tuning, we do posit that it is physiologically plausible for a downstream population to make use of these key prior distributions in implementing an adaptive decoding strategy: the same brain regions across individuals of a species encode similar information using similar strategies and experience drift with similar statistics. Readouts across populations that connect regions therefore have access to reasonable priors about their upstream neural activity, which may be sufficient to implement adaptive decoding strategies.

We conclude with a call for greater scrutiny of the statistics of drift across brain regions. While drift is necessarily a gradual process at the scale of populations, this drift still permits a variety of underlying processes operating at the level of individual neurons. We predict that the neural populations with the highest rates of drift and the highest rates of plasticity, changes in representation will take place through neurons

performing sudden jumps away from previous tuning values in order to avoid ambiguity in downstream readouts.

Methods

Adaptive Decoding and Drift Simulations

Solving the optimisation for the gradual drift model

We use gradient ascent to solve for the most likely values of $\widehat{\boldsymbol{\theta}}_t$.

$$\mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}_t) = \sum_i^N \sum_j^N \log h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) + \sum_i^N \log f(\theta_i, \widehat{\theta}_{i,t-1}) \tag{22}$$

We note that the left summation depends only on the neuron behaviour model and the right summation depends only on the drift model. In this simple example, we can find analytical forms of the Jacobian and the Hessian, which makes solving via gradient ascent faster. This makes use of two important properties of h. First, it is symmetrical over i,j:

$$h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) \equiv h(\mathbf{x}_j, \mathbf{x}_i, \theta_j, \theta_i) \tag{23}$$

Second, in the special case when $i = j$, all choices of θ_i are equivalent.

$$\frac{\partial}{\partial \theta_i} \left(h(\mathbf{x}_i, \mathbf{x}_i, \theta_i, \theta_i) \right) = 0 \tag{24}$$

Combining these properties and taking the derivative of the log-likelihood with respect to a single parameter:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 2 \sum_{j \neq i}^N \left(\frac{\partial}{\partial \theta_i} \left(\log h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) \right) \right) + \frac{\partial}{\partial \theta_i} \left(\log f(\theta_i, \widehat{\theta}_{i,t-1}) \right) \tag{25}$$

And then, to produce the Hessian:

$$\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} = \begin{cases} \frac{\partial^2}{\partial \theta_i^2} \left(\log f(\theta_i, \widehat{\theta}_{i,t-1}) \right) & \text{if } i = j \\ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left(\log h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j) \right) & \text{otherwise} \end{cases} \tag{26}$$

We initialise of the parameters for the gradient ascent is to use their values from the previous day, $\widehat{\boldsymbol{\theta}}_{t-1}$. A complete expression and its derivation is provided in the supplementary material.

Exact solution to the optimisation for the sudden drift model

The discontinuity in f_{sudden} means that it is ill-conditioned for a gradient ascent-based solution applied to the entire parameter space. Instead, we consider a solution that decomposes the problem into smaller components that can be individually solved by gradient ascent (and, in the subsequent section, a technique for approximating this solution for large populations).

For each neuron, there are two possible cases: either the neuron keeps its tuning parameter, or it retunes. In the first case, trivially $\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1}$. In the second case, the neuron needs a tuning parameter update *and*

should be withheld from use in estimating the parameters of other neurons this day, as $\widehat{\boldsymbol{\theta}}_{t-1}$ is no longer accurate. To solve, we perform two steps:

- Infer which neurons have kept their tuning parameter intact
- Use this subset of neurons to adapt the parameters of the remaining neurons

Let the set A contain the neurons which have retained their tuning parameters. A neuron i is a member of this set if and only if $u_i = 1$, where:

$$u_i(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{t-1}) = \begin{cases} 1 \text{ if } P(\mathbf{X}|\text{changed})P(\text{changed}) < P(\mathbf{X}|\text{changed}')P(\text{changed}') \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

In the concrete example for the sudden drift model, the above condition can be written as:

$$\int_0^{2\pi} g(\mathbf{x}_i, \psi, \widehat{\boldsymbol{\theta}}_{t-1}) \frac{\alpha}{2\pi} \mathrm{d}\psi < g(\mathbf{x}_i, \widehat{\theta}_{i,t-1}, \widehat{\boldsymbol{\theta}}_{t-1})(1-\alpha) \tag{28}$$

Where $g(\mathbf{x}_i, \psi, \widehat{\boldsymbol{\theta}}_{t-1})$ is the probability of the observations $\mathbf{x}_i$ given a tuning angle ψ.

This allows us to express the update rule as:

$$\widehat{\boldsymbol{\theta}}_t = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}) \simeq \begin{bmatrix} z_0(\theta_0, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}) \\ \dots \\ z_i(\theta_i, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}) \\ \dots \\ z_N(\theta_N, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}) \end{bmatrix} \tag{29}$$

Where:

$$z_i(\theta_i, \widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{X}) = \begin{cases} \widehat{\theta}_{i,t-1} & \text{if } i \in A \\ \arg\max_{\boldsymbol{\theta}} \sum_{j \in A} \log h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \widehat{\theta}_{j,t-1}) & \text{otherwise} \end{cases} \tag{30}$$

Noting that the use of the approximately equal symbol, as this solution discards the small amount of information available in the pairwise mutual tuning between neurons not in A.

Approximate solution to the optimisation for the sudden drift model

In practice, evaluating the condition for u_i becomes computationally intractible for large M due to resolution limitations of floating point number representations: it requires the product of very many small numbers. Circumventing this limitation in the typical fashion–a sum of logarithms–requires a closed form of the integral of g, which is not necessarily available. We instead choose to make an approximation in determining set membership of A.

We evaluate $\log g(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{t-1})$ for each neuron in the population and are therefore able to rank each neuron in the population by conformity to expectation. For a drift rate parameterised by α and a number of neurons in the population N, the number of neurons that drifted is binomially distributed. We can select a probability threshold q (by using the CDF of the binomial distribution) that results in us excluding the bottom Q neurons from A by ranking.

The choice of q is a hyperparameter. If q is large (e.g. 0.95), many of the neurons excluded from A are likely to not require retuning. For small N and high drift rates (large α), this can exclude a sizeable portion of the population and reduce the accuracy of tuning parameter estimates. Conversely, if q is small (e.g. 0.5), neurons that require re-estimating will be included in A and not have their parameters updated. In all simulations shown in this work, we set $q = 0.9$.

We note that this approximation has the potential to disadvantage an adaptive decoder compensating for sudden drift when compared to one compensating for gradual drift. Indeed, we observe that levels of noise high enough to make an adaptive decoder perform no better than a fixed decoder, this approximation leads to performance that is worse than rather than equivalent to a fixed decoder (as visible in Fig. 2.e and Fig. S2.c). However, given our choice of neuron population sizes *and* the continued outperformance of the sudden drift adaptive decoders, we take our use of this approximation to be immaterial to our conclusions.

Implementation of solvers

The gradient ascent components of these solutions were implemented using the `scipy`'s library's `optimize.minimize` function (in the case of the multivariate optimisation for the gradual case) and the `optimize.minimize_scalar` function (in the case of the independent scalar optimisations for the sudden case) [26]. We note the additional following details:

- Solution bounds are set to -2π to 2π, and only remapped to the range $-\pi$ to π after convergence
- The underlying solver method is based on `L-BFGS-B` [27]
- We use the analytical expression for the Jacobian as described in the supplementary material

Tuning parameterisation-free drift characterisation

In-vivo distributions of correlation changes

For in-vivo data, 'correlation' referes to the Pearson Correlation Coefficient. When visualising the distribution of correlation changes for experiments T sessions apart (in days for the simulated and Driscoll *et al.* (2017) dataset, in weeks for the Marks & Goard (2021) dataset), we include all session pairs separated by that interval. For example, if $T = 2$ we would include day 1 paired with day 3, day 2 paired with day 4, day 3 paired with day 5, and so on. Correlations between neuron pairs can only be evaluated within an individual animal. We aggregate all neuron pair correlations from all animals to produce the distributions. For both in-vivo datasets, we select $C_{\min} = 0.5$: this is high enough to cull spurious correlations (which mask the visibility of the effect of drift on the distribution), while low enough to include many pairs of neurons.

Mixture of Gaussian fits

We fit the distribution of correlation changes with a 2-component mixture of Gaussians. This is solved using the EM algorithm as implemented by the `scikit-learn` library [28]. For the in-vivo data, we use bootstrapping (100 iterations per distribution) to fit many resampled distributions to produce the SEM used for the error-bars in Fig. 3.g and Fig. 3.h.

Modelling the movement of the upper mean of the mixture of Gaussians fit

In the case of sudden drift, we approximate fits to distribution of correlation changes as always being bimodal with a fixed upper mean (i.e. it is time independent).

$$\mu_{2,\text{sudden}}(t) = c \tag{31}$$

In the case of gradual drift, we approximate the upper mean of the mixture model fit to increase as a function of elapsed time, described by:

$$\mu_{2,\text{gradual}}(t) = c\,(1 - e^{-bt}) \tag{32}$$

For the in-vivo data, we fit a weighted mixture of the sudden and gradual models:

$$\mu_{2,\text{composite}}(t) = a\mu_{2,\text{gradual}}(t) + (1-a)\mu_{2,\text{sudden}}(t) \tag{33}$$

Noting the following constraints:

$$0 \leq a \leq 1 \tag{34}$$

$$b \geq 0 \tag{35}$$

$$0 \leq c \leq 1 \tag{36}$$

The parameters a, b, and c are found by minimising the mean squared error between this model of upper mean locations and its measured values.

Drift in the posterior parietal cortex

Data Inclusion

The Driscoll *et al.* (2017) dataset comprises recordings from 5 animals (m01 through m05). Because our analysis relies on pairs of neurons being visible across sessions, we note that the number of individual neurons strongly informs the quality of the data. Our analysis includes only neurons that are identified between sessions at the highest level of confidence as labelled in the dataset. Animals m02 and m05 have a distinctly lower total number of neurons identifiable between sessions (Fig. S3.a), experience a more rapid decline in the number of neurons identifiable between sessions as the number of days between those sessions increase, and exhibit a drop off over time in the number of identifiable neurons. We note that consequently, while we include all animals in our analysis, that animals m01, m03, and m04 are more represented in the aggregated data, but the proportion of representation does not change over time (Fig. S3.b).

Data Processing

To calculate the activity correlation between neuron pairs during each session, we first convolved spike times (from the original dataset as extracted by deconvolving Calcium traces) with a Gaussian window (standard deviation equivalent to 0.45 seconds). We then evaluated pairwise temporal correlation between these traces.

We did not exclude neural activity during the interval of darkness between trials.

Drift in the visual cortex

Data Inclusion

The Marks & Goard *et al.* (2021) dataset comprises recordings from 12 animals. Mice 7 and 8 are omitted from our analysis as their imaging schedules deviate from the one used by the other animals. Mouse 2 is omitted from our analysis as, uniquely, its surgery provides two separate imaging windows. Our analysis includes only neurons labelled as having a registration quality of 3 or higher (unambiguously identified across sections). Our analysis makes use of only the experimental protocol which exposes animals to the movies of naturalistic scenes.

Data Processing

To calculate the activity correlation between neuron pairs during each session, we convolved the raw Calcium traces with a Gaussian window (standard deviation equivalent to 0.25 seconds). These traces were subsequently used as the basis for pairwise temporal correlation.

References

[1] Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147, October 2019.

[2] Michael E Rule, Adrianna R Loback, Dhruva V Raman, Laura N Driscoll, Christopher D Harvey, and Timothy O'Leary. Stable task information from an unstable neural population. *eLife*, 9:e51121, July 2020.

[3] Michael E. Rule and Timothy O'Leary. Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences*, 119(7):e2106692119, February 2022. Publisher: Proceedings of the National Academy of Sciences.

[4] Laura N. Driscoll, Lea Duncker, and Christopher D. Harvey. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, October 2022.

[5] Charles Micou and Timothy O'Leary. Representational drift as a window into neural and behavioural plasticity. *Current Opinion in Neurobiology*, 81:102746, August 2023.

[6] Dhruva V Raman and Timothy O'Leary. Optimal plasticity for memory maintenance during ongoing synaptic change. *eLife*, 10:e62912, September 2021.

[7] Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M. Sengupta, Dmitri B. Chklovskii, and Cengiz Pehlevan. Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nature Neuroscience*, 26(2):339–349, February 2023.

[8] Pratik Chaudhari and Stefano Soatto. Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks. In *IEEE Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, February 2018.

[9] Kyle Aitken, Marina Garrett, Shawn Olsen, and Stefan Mihalas. The geometry of representational drift in natural and artificial neural networks. *PLOS Computational Biology*, 18(11):e1010716, November 2022.

[10] Carl E. Schoonover, Sarah N. Ohashi, Richard Axel, and Andrew J. P. Fink. Representational drift in primary olfactory cortex. *Nature*, 594(7864):541–546, June 2021.

[11] Farhad Pashakhanloo and Alexei Koulakov. Stochastic Gradient Descent-induced drift of representation in a two-layer neural network, February 2023. arXiv:2302.02563 [cond-mat].

[12] Yaniv Ziv, Laurie D. Burns, Eric D. Cocker, Elizabeth O. Hamel, Kunal K. Ghosh, Lacey J. Kitch, Abbas El Gamal, and Mark J. Schnitzer. Long-term dynamics of CA1 hippocampal place codes. *Nature Neuroscience*, 16(3):264–266, March 2013.

[13] Dorgham Khatib, Aviv Ratzon, Mariell Sellevoll, Omri Barak, Genela Morris, and Dori Derdikman. Experience, not time, determines representational drift in the hippocampus. *bioRxiv*, pages 2022–08, 2022.

[14] Nitzan Geva, Daniel Deitch, Alon Rubin, and Yaniv Ziv. Time and experience differentially affect distinct aspects of hippocampal representational drift. *Neuron*, 111(15):2357–2366.e5, August 2023. Publisher: Elsevier.

[15] Daniel Deitch, Alon Rubin, and Yaniv Ziv. Representational drift in the mouse visual cortex. *Current Biology*, 31(19):4327–4339.e6, October 2021.

[16] Tyler D. Marks and Michael J. Goard. Stimulus-dependent representational drift in primary visual cortex. *Nature Communications*, 12(1):5169, August 2021.

[17] Laura N. Driscoll, Noah L. Pettit, Matthias Minderer, Selmaan N. Chettih, and Christopher D. Harvey. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell*, 170(5):986–999.e16, August 2017.

[18] Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021. Number: 7858 Publisher: Nature Publishing Group.

[19] William A. Liberti, Tobias A. Schmid, Angelo Forli, Madeleine Snyder, and Michael M. Yartsev. A stable hippocampal code in freely flying bats. *Nature*, 604(7904):98–103, April 2022.

[20] Sadra Sadeh and Claudia Clopath. Contribution of behavioural variability to representational drift. *eLife*, 11:e77907, August 2022.

[21] Noah L. Pettit, Xintong C. Yuan, and Christopher D. Harvey. Hippocampal place codes are gated by behavioral engagement. *Nature Neuroscience*, 25(5):561–566, May 2022.

[22] Pamela J. Kennedy and Matthew L. Shapiro. Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *Proceedings of the National Academy of Sciences*, 106(26):10805–10810, June 2009.

[23] Christopher D. Harvey, Philip Coen, and David W. Tank. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–68, April 2012. Publisher: Nature Publishing Group.

[24] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2L: Contrastive Continual Learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9516–9525, 2021.

[25] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. Publisher: Proceedings of the National Academy of Sciences.

[26] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[27] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, August 1989.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Supplements

Derivations

Expected correlation between neuron pairs

$$E[\mathbf{x_i x_j}] = \int_{r=0}^{r=1} \int_{\phi=-\pi}^{\phi=\pi} p_r(r) p_\phi(\phi) a(r,\phi,0) a(r,\phi,y) \mathrm{d}\phi \mathrm{d}r \tag{37}$$

We note that $a(r,\phi,\theta)$ is zero for $|\phi-\theta| > \pi/2$. Therefore, taking $0 \leq y < \pi$:

$$E[\mathbf{x_i x_j}] = \int_{r=0}^{r=1} \int_{\phi=y-\pi/2}^{\phi=\pi/2} p_r(r) p_\phi(\phi) a(r,\phi,0) a(r,\phi,y) \mathrm{d}\phi \mathrm{d}r \tag{38}$$

$$= \int_{r=0}^{r=1} \int_{\phi=y-\pi/2}^{\phi=\pi/2} p_r(r) p_\phi(\phi) r^2 \cos^2(\phi) \cos^2(\phi-y) \mathrm{d}\phi \mathrm{d}r \tag{39}$$

Noting that r and ϕ follow uniform distributions:

$$E[\mathbf{x_i x_j}] = \int_{r=0}^{r=1} \int_{\phi=y-\pi/2}^{\phi=\pi/2} \frac{1}{1}\frac{1}{2\pi} r^2 \cos^2(\phi) \cos^2(\phi-y) \mathrm{d}\phi \mathrm{d}r \tag{40}$$

$$= \frac{1}{6\pi} \int_{\phi=y-\pi/2}^{\phi=\pi/2} \cos^2(\phi) \cos^2(\phi-y) \mathrm{d}\phi \tag{41}$$

$$= \frac{1}{6\pi}\frac{1}{16} \left(3\sin 2y + 2(\pi-y)(\cos 2y + 2)\right) \tag{42}$$

Jacobian for gradient ascent in gradual drift model

Then, the first partial derivative of $\log h$:

$$\frac{\partial}{\partial\theta_i} \left(\log h(\mathbf{x}_i, \mathbf{x}_j, \theta_i, \theta_j)\right) = \frac{M}{\sigma^2} \left(\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{M} - \mu(|\theta_i - \theta_j|)\right) \frac{\partial}{\partial\theta_i} \left(\mu(|\theta_i - \theta_j|)\right) \tag{43}$$

Where:

$$\frac{\partial}{\partial\theta_i} \left(\mu(|\theta_i - \theta_j|)\right) = \frac{1}{24\pi} \frac{\theta_i - \theta_j}{|\theta_j - \theta_i|} \left(\cos(2|\theta_i - \theta_j|) + (|\theta_i - \theta_j| - \pi)\sin(2|\theta_i - \theta_j|) - 1\right) \tag{44}$$

For the gradual drift prior, we have:

$$\sum_i^N \log f(\theta_i, \widehat{\theta}_{i,t-1}) = N \log \frac{1}{\rho\sqrt{2\pi}} - \frac{1}{2\rho^2} \sum_i^N (\theta_i - \widehat{\theta}_{i,t-1})^2 \tag{45}$$

And the partial derivative:

$$\frac{\partial}{\partial\theta_i} \left(\log f(\theta_i, \widehat{\theta}_{i,t-1})\right) = \frac{\widehat{\theta}_{i,t-1} - \theta_i}{\rho^2} \tag{46}$$

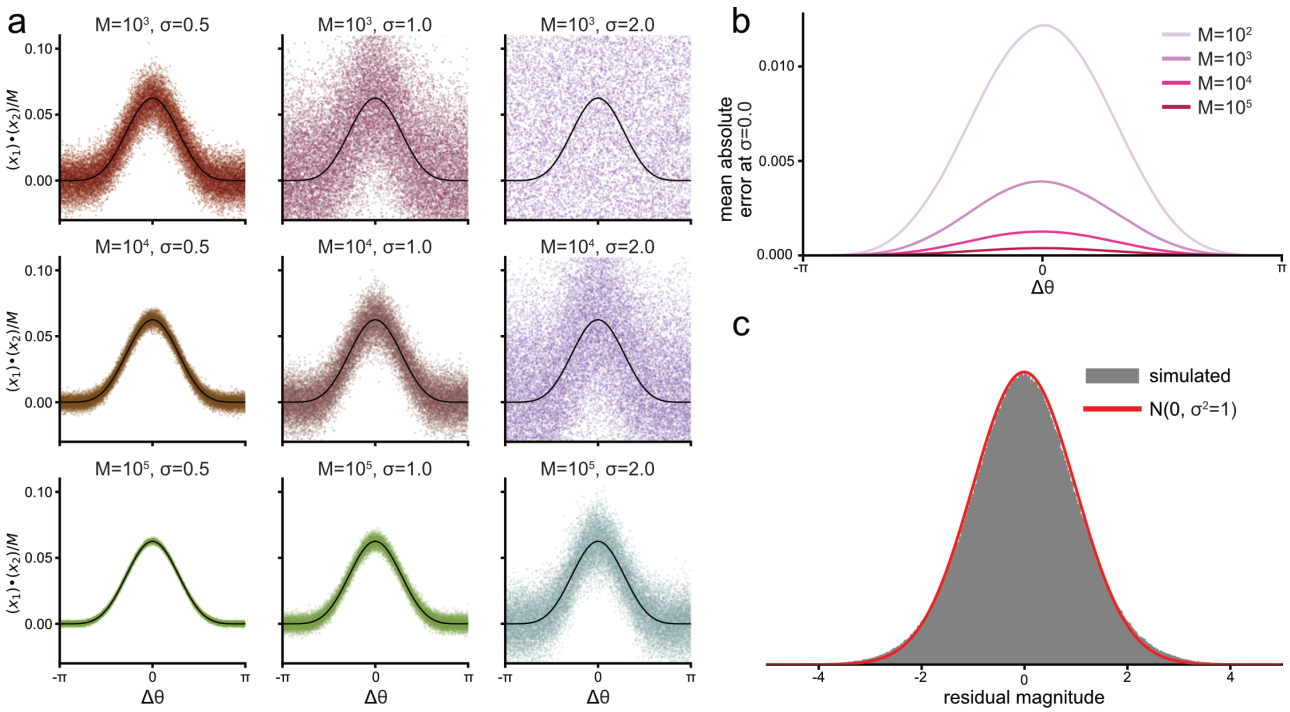

Fig. S1: **(a)** Visualisations of the distribution of correlations between neuron pairs separated by an angular distance $\Delta\theta$ for a variety of observation noises σ and total numbers of samples M. The black trace shows approximation for the expected value of the correlation. **(b)** The mean absolute error between the approximation for the expected value of the pairwise correlation and its actual value as computed by a Monte Carlo simulation, shown for in the absence of any observation noise ($\sigma = 0$) as a function of angular distance between the two neurons. **(c)** Distribution of the error residual, the difference between the expected value of the pairwise correlation and its actual value, in the presence of observation noise ($\sigma = 1.0$), presented with a comparison to the normal distribution, simulated at $M = 10^3$.

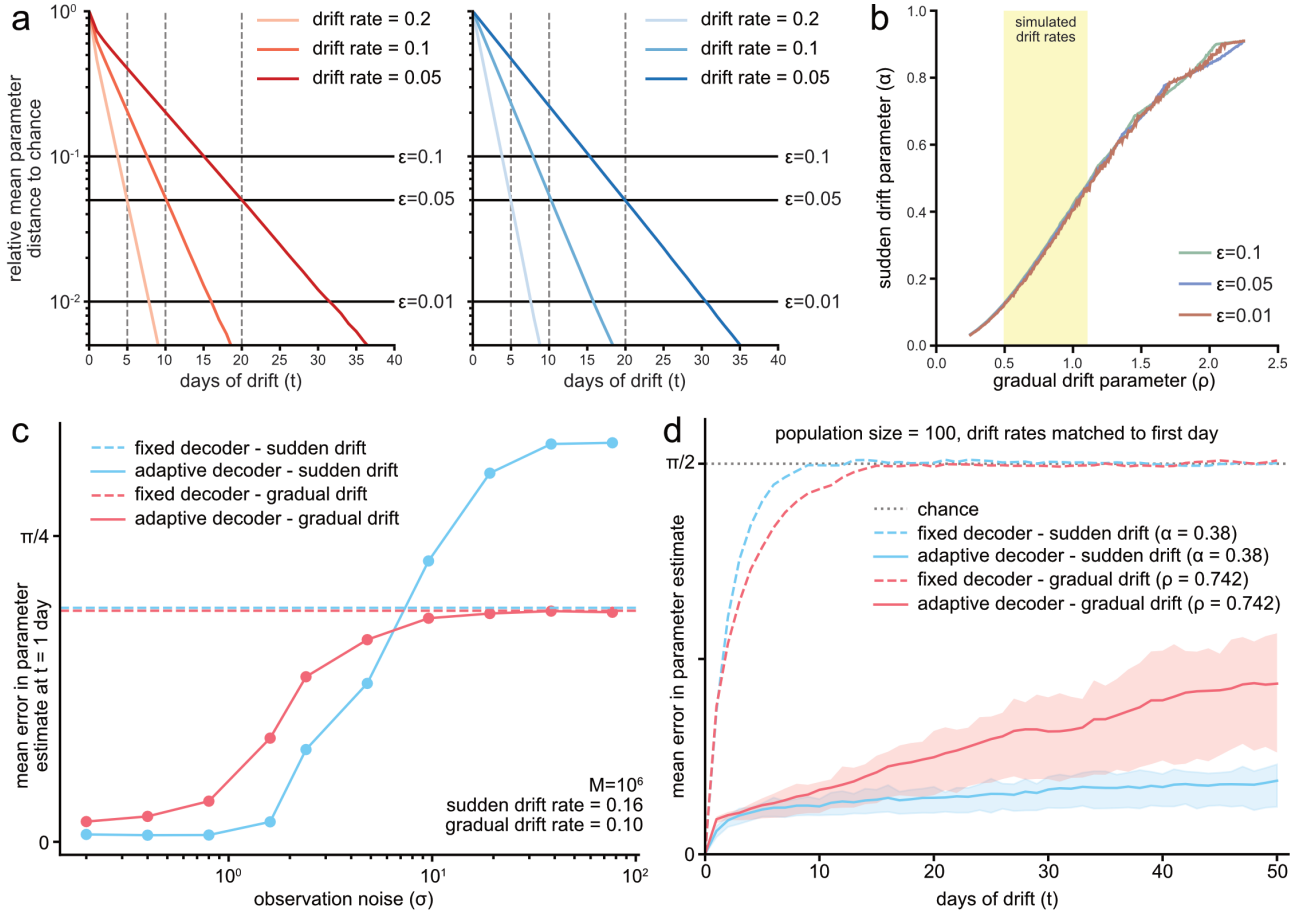

Fig. S2: **(a)** The approach of tuning parameters to their chance values over the course of several days of drift, mean of 10^6 simulations. Shown for the gradual drift model (left, red) and the sudden drift model (right, blue). Drift rates in the legend listed for $\epsilon = 0.05$. **(b)** The equivalence mapping between the gradual and sudden drift parameters for multiple values of ϵ, highlighting the range of gradual drift rates used in the simulations of this article. **(c)** Error in tuning parameter estimates after the first day of drift as a function of the observation noise σ, using alternative equivalence between gradual and sudden drift that induces the same amount of error in a fixed decoder after the first day of drift. **(d)** As in Fig. 2.f, but using the alternative drift equivalence.

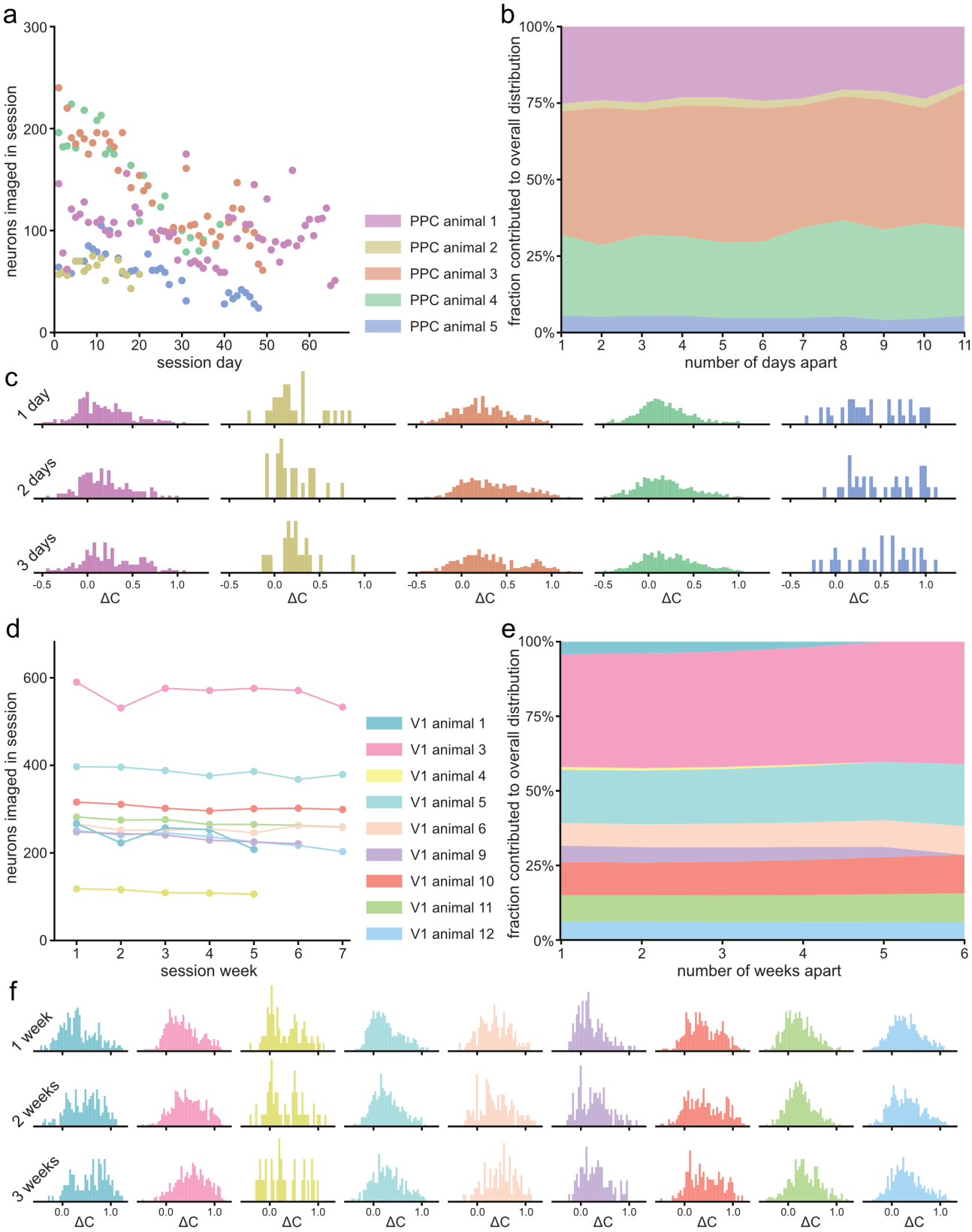

Fig. S3: **(a)** The number of regions of interest (ROIs) labelled as neurons in each recording session for each animal of the Driscoll *et al.* PPC dataset. **(b)** The proportional contribution of each animal to the aggregated distribution of scaled changes in pairwise correlation. **(c)** Distributions of changes in pairwise correlation for the individual animals, shown for 1, 2, and 3 sessions apart. **(d)** The number of ROIs labelled as neurons in each recording session for each animal of the Marks & Goard V1 dataset. **(e)** As in (b), shown for the V1 dataset. **(f)** As in (c), shown for the V1 dataset.